Evaluating Large Language Models on Slovak Bebras Tasks: A Multimodal Computational Thinking Benchmark

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly applied to reasoning tasks, yet their behaviour on less-resourced languages and on multimodal computational thinking problems is still poorly understood. We introduce a dataset of **822 tasks** collected from the Slovak *iBobor* competition, the national instance of the international Bebras Challenge, and use it to evaluate six contemporary LLMs (GPT-5.2, GPT-5.4 Mini, Gemini 2.5 Pro, Gemini 2.5 Flash, Llama 4 Maverick, Llama 4 Scout) accessed via OpenRouter. We report accuracy, runtime, and token usage across three input modalities (text-only, single-image, multi-image), across competition categories (including the text-only *Nevidiaci* categories), and against average student performance for three competition years. We additionally compare two input representations: structured text-with-images and full-task screenshots. We find that (i) the strongest model (GPT-5.2) matches or exceeds the published average student accuracy in most categories, while smaller models lag behind, (ii) accuracy is associated far more with input modality than with nominal task difficulty (it degrades sharply with the number of images but only weakly across age-based categories), and (iii) rendering short tasks as screenshots increases token usage several-fold, with mixed and confounded effects on accuracy, contrary to recent findings on long textual inputs. The student comparison is category-level rather than task-matched (see Limitations).

1 Introduction

Large language models (LLMs) are now routinely applied to tasks that go beyond text generation, including multi-step reasoning, code synthesis, and visual question answering. Their reasoning capabilities, however, remain contested: models that produce fluent prose can still fail at problems that humans solve reliably, and the gap is amplified for less-resourced languages and for inputs that mix text and images (Huang et al., 2023; Salazar et al., 2025).

The *Bebras International Challenge on Informatics and Computational Thinking* (Bebras International Challenge on Informatics and Computational Thinking) is well suited as an evaluation testbed. Designed for school students with little or no programming background, it consists of short, self-contained, multiple-choice problems that target logic, abstraction, pattern recognition, and algorithmic reasoning. The tasks are not synthetic benchmarks written with AI in mind; they are produced annually by educators across more than 50 countries, are vetted for human-solvability, and have well-defined correct answers. In Slovakia, the contest runs under the name *iBobor* (iBobor), providing a natural source of evaluation material in a Slavic language that is largely absent from major multilingual benchmarks.

The most directly comparable prior evaluation of LLMs on Bebras-style problems (Capdehourat et al., 2025) uses English (or relies on translation) and either includes only a small set of tasks or strips out visual content, while broader native multilingual multimodal exam benchmarks (Salazar et al., 2025) omit Slovak. We close both gaps by assembling a large Slovak *iBobor* dataset and evaluating modern LLMs natively on it, preserving both textual and visual content.

Research questions. We ask: (RQ1) *How accurately do current LLMs solve Slovak iBobor tasks across task types and competition categories?* (RQ2) *How does input modality (text only, one image, multiple images) affect performance?* (RQ3) *How does model performance compare to published average human (student) performance for the same competition categories?* (RQ4) *How does input format (structured text-with-images vs. a single screenshot) influence token usage and accuracy?*

Contributions.

- An **822-task Slovak Bebras dataset** covering 16 competition years and 11 categories, with text and image content preserved in a machine-readable JSON structure that supports tasks with multiple images.
- A controlled **evaluation of six LLMs** from three providers (OpenAI, Google, Meta), reporting accuracy, runtime, and token usage under a unified API.
- Analyses by **task type, competition category, and against student performance**, showing that accuracy varies more with input modality than with age-based category difficulty, and that the strongest model matches or exceeds the published average student in most categories.
- A **token-efficiency comparison** between structured and screenshot-based input representations, with implications for short-task multimodal evaluation.

2 Background and Related Work

Bebras, iBobor, and computational thinking. Computational thinking, the set of problem-solving skills associated with computer science such as abstraction, decomposition, pattern recognition, and algorithmic reasoning, has been a target of education research since Wing (2006). The *Bebras* challenge operationalises it for school students through short, self-contained tasks that are deliberately designed to probe these skills without requiring programming, following published criteria for task quality and skill coverage (Dagiene and Futschek, 2008; Bebras International Challenge on Informatics and Computational Thinking). In Slovakia the contest runs as *iBobor* (iBobor), with age-based categories (*Drobec*, *Bobrík*, *Benjamín*, *Kadet*, *Junior*, *Senior*) plus dedicated *Nevidiaci* categories for visually impaired students whose tasks are purely textual (*Nevidiaci ZŠ/SŠ*: primary/secondary school). Tasks are four-option multiple-choice and combine text with images, diagrams, or interactive elements. This educator-authored, human-vetted provenance is what makes Bebras attractive as an evaluation source: the items are not written with language models in mind.

Multimodal LLMs and reasoning benchmarks. Vision-language models that interleave images and text (Alayrac et al., 2022; Liu et al., 2023; OpenAI, 2023; Gemini Team, Google, 2023) are now routinely evaluated on benchmarks that demand visual reasoning rather than surface text matching: MMMU (Yue et al., 2024), MATHVISTA (Lu et al., 2024), SCIENCEQA (Lu et al., 2022), and MMBENCH (Liu et al., 2024). A recurring finding is that models degrade sharply when a task genuinely requires perceiving and integrating visual content; benchmarks targeting *multi-image* reasoning specifically, such as BLINK (Fu et al., 2024) and MUIRBENCH (Wang et al., 2025), report large drops. This is directly relevant to our central result: *iBobor* accuracy falls monotonically from text to single-image to multi-image tasks.

Multilingual and multilingual–multimodal evaluation. Text-only multilingual suites such as BELEBELE (Bandarkar et al., 2024), MEGA (Ahuja et al., 2023), AYA (Üstün et al., 2024), and GLOBAL-MMLU (Singh et al., 2025) document large, uneven cross-lingual gaps; BELEBELE in fact includes Slovak, but only as text reading comprehension. Native (non-translated) multilingual *multimodal* evaluation is far rarer: M3EXAM (Zhang et al., 2023), CVQA (Romero et al., 2024), and KALEIDOSCOPE (Salazar et al., 2025) are the closest. Salazar et al. (2025) assemble 20,911 in-language multiple-choice questions across 14 subjects and 18 languages and report that all systems lose several points on multimodal relative to text-only items and that less-resourced languages are underrepresented; Slovak is not among their 18 languages. Our benchmark adds native Slovak, genuine (and frequently multi-image) multimodality, educator-authored computational-thinking items, and a per-category human reference.

Slovak and low-resource NLP. Existing Slovak NLP resources, such as SLOVAKBERT (Pikuliak et al., 2022), are predominantly text-only, as are the multilingual suites discussed above. None evaluate multimodal computational-thinking reasoning, the gap *iBobor* fills; we therefore position our resource as complementary to language-understanding benchmarks rather than as a general Slovak NLU resource.

LLMs on Bebras. The closest related work is Capdehourat et al. (2025), who evaluate GPT-4o, GPT-4o-mini, Gemini, and Llama variants on 26 Bebras tasks adapted from the Uruguayan competition, with images converted to text descriptions. They report GPT-4o reaching ≈70% accuracy on

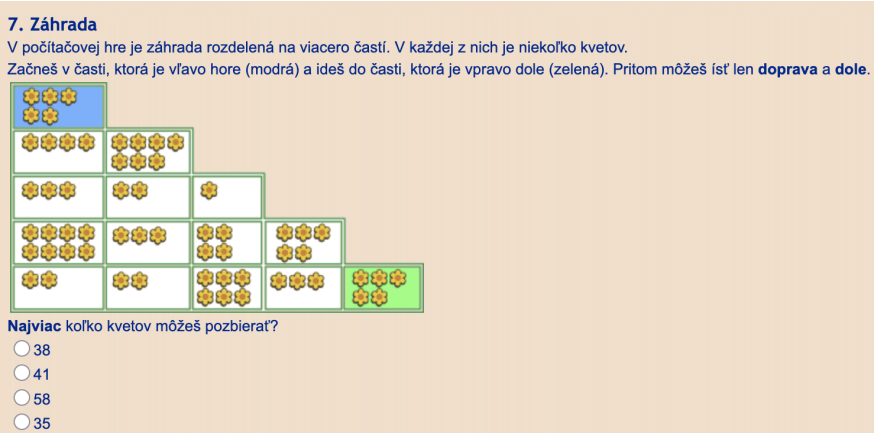

Figure 1: Example of an image-based iBobor task, *Záhrada* (*Garden*), Senior category, 2019/2020. The visual component is required to determine the correct answer.

some sets, while smaller models stay around 40–50%. Compared to that work, we (i) keep the images as images, (ii) work in Slovak instead of English, and (iii) scale up from 26 to 822 tasks, large enough for per-modality and per-category breakdowns.

Token efficiency of text-as-image inputs. Li et al. (2025) show that, for long-context retrieval and summarisation, rendering text as images can roughly halve decoding tokens with little accuracy loss. *iBobor* tasks are short, so it is unclear whether the same trade-off applies; we test this directly in Section 5.5.

Benchmark contamination. Because the *iBobor* archive has been public since 2007, memorisation is a genuine risk for any model trained on web text. A growing literature documents and detects such contamination (Sainz et al., 2023; Golchin and Surdeanu, 2024). We do not run a dedicated detection probe; our per-year check (Section 5.6) is only a weak indirect signal and does not rule out memorisation of older tasks, as we make explicit there.

3 Dataset

3.1 Source and Collection

We collected tasks from the public demo of the Slovak *iBobor* platform (iBobor), with the consent of the national coordinator of the Slovak *iBobor* competition. We restricted the dataset to publicly accessible multiple-choice tasks, since open-ended or fully interactive tasks cannot be mapped to a single correct answer in an automatically evaluable manner.

Because tasks are served through an interactive web interface rather than as downloadable files, we implemented a Python scraping pipeline based on the `Crawlee` (Apify) `PlaywrightCrawler`. For every available year and category, the pipeline iterates through all task pages and extracts the title, statement, answer choices, associated images, year, category, and task type, plus the correct answer from the post-competition results page. Selected HTML elements (tables, bold, italic) are preserved, since some tasks rely on formatting for correctness. Each task additionally has a full-page screenshot stored separately for the screenshot-based experiment in Section 5.5.

3.2 Task Types

During extraction, tasks were grouped into four content-structure types: **text** (statement and options are pure text; see Figure 3); **image** (one or more images form part of statement or options and are necessary for correct interpretation; see Figure 1); **canvas** (interactive tasks requiring on-screen manipulation); and **Flash** (legacy tasks incompatible with modern browsers, present mostly in 2016/17 and earlier). Canvas and Flash tasks were excluded from the final dataset because they do not follow the multiple-choice format.

3.3 Representation

Each retained task is stored as a single JSON object containing an integer `id`, the task `title` and `question` text, an array of `choices` (each a `value`/`label` object), an `images` array, the `year`, `category`, detected `type`, and the correct answer (as an index and as plain text). Images embedded in the statement or in option labels are

replaced by indexed placeholders (`<image_1>`, `<image_2>`, ...) referring into the `images` array. This extends the Kaleidoscope-style format with support for multiple images per task, which is essential because slightly more than half of our image-type tasks (321 of 629) contain more than one image (Table 4).

3.4 Statistics

The pipeline collected 1098 tasks; after filtering out canvas and Flash tasks, 822 remain. Of these, 193 are text-only and 629 are image-based, with 308 containing a single image and 321 containing multiple images. Every retained task has exactly four answer options, so the random-choice baseline is exactly 0.25. The dataset spans 16 competition years and 11 categories, including three non-standard categories (*3.–4. roč.*, *5.–7. roč.*, *6.–9. roč.*) used only in the special years *Karanténa* and *Prechádzky*, and the two *Nevidiaci* categories whose tasks are purely textual. The dataset is an evaluation-only benchmark: it has no train/development/test split, and every task is used exactly once for evaluation. Table 4 reports the full counts.

4 Experimental Setup

4.1 Models

We evaluate six models from three providers, including, where applicable, both a stronger and a more efficient variant per family: `openai/gpt-5.2`, `openai/gpt-5.4-mini`, `google/gemini-2.5-pro`, `google/gemini-2.5-flash`, `meta-llama/llama-4-maverick`, and `meta-llama/llama-4-scout`. All models are accessed through the unified OpenRouter API (OpenRouter), so request handling, response parsing, and error reporting are identical across providers. Experiments were run between January and March 2026 with the model versions then available. We do not train or fine-tune any model and use no local GPU compute: every model is queried as a hosted, third-party API. Exact parameter counts are not publicly disclosed for the proprietary GPT and Gemini models; the open-weight Llama 4 models are mixture-of-experts systems with 17B active parameters each (Maverick: 400B total, 128 experts; Scout: 109B total, 16 experts). As every model is queried via a hosted API, in place of GPU-hours we report the total wall-clock runtime and mean per-task token usage of each run (Table 1) as our compute-cost proxy.

4.2 Evaluation Pipeline

Tasks are split by modality into three subsets (text-only, single-image, multi-image), each handled by a dedicated script that submits tasks sequentially to a model, parses the response, and logs both the prediction and the token usage reported by the API. Each task is evaluated **once** per model, with no majority voting; this mirrors the single-attempt setting of the real *iBobor* competition. A separate supplementary pipeline submits each task as a full-page screenshot for the comparison in Section 5.5.

4.3 Prompting

We use a fixed zero-shot prompt for every model. The system message describes the task format and required JSON output; the exact text-only system message is reproduced verbatim in Appendix A. For multimodal inputs, the system message is extended to declare that images form part of the question and, in the multi-image case, that answer options may also contain images. Image placeholders in the task text are replaced by the corresponding images in order, and images are passed as base64 data URLs in the multimodal request. We set `temperature=0` for reproducibility and pass `response_format` with type `json_object` where supported. We do not set `max_tokens`, since one of our outcomes is the comparison of actual token usage across input representations. We perform no hyperparameter or prompt search: a single fixed zero-shot configuration is used for every model and setting.

4.4 Scoring

Returned `choice_value` fields are compared against the gold answer value. If a model returns a label instead of a value, the label is mapped back to a value where unambiguous; otherwise the response is logged as an error. For the screenshot setting, the returned answer text is compared to the plain-text gold answer after lightweight normalisation (Appendix A). We report accuracy, total runtime, and average tokens per task (prompt + completion).

All reported accuracies are point estimates from a single run with `temperature`=0. To quantify sampling uncertainty we treat each accuracy as a

Model	Corr.	Acc.	Time	Tok/q
GPT-5.4 Mini	307	0.374	14m	489
GPT-5.2	647	**0.787**	9h 48m	2847
Gemini 2.5 Flash	357	0.434	21m	1153
Gemini 2.5 Pro	516	0.628	12h 58m	6729
Llama 4 Maverick	335	0.408	16m	885
Llama 4 Scout	284	0.346	11m	991

Table 1: Overall results on all 822 tasks. **Corr.** is correctly answered tasks; **Tok/q** is mean total tokens (prompt + completion) per task. Random-baseline accuracy is 0.25. 95% Wilson confidence intervals on **Acc.** are $\approx \pm 0.03$ (n=822), so GPT-5.2's lead over Gemini 2.5 Pro and the gaps to the remaining models are well outside sampling noise.

Model	Text (n=193)	1-img (n=308)	Multi-img (n=321)
GPT-5.4 Mini	0.523	0.364	0.293
GPT-5.2	**0.974**	**0.831**	**0.632**
Gemini 2.5 Flash	0.617	0.419	0.340
Gemini 2.5 Pro	0.938	0.656	0.414
Llama 4 Maverick	0.591	0.367	0.336
Llama 4 Scout	0.446	0.331	0.299

Table 2: Accuracy by input modality (the three-way text / single-image / multi-image split). The same monotone drop is visible for every model: text > one image > multiple images. The multi-image column uses all 321 tasks; cf. Table 5, whose multi-image cell is the 111-task textual-option subset.

binomial proportion and report 95% Wilson confidence intervals: these are $\approx \pm 0.03$ for the overall numbers (n=822; Table 1), roughly ± 0.02 to ± 0.07 for the per-modality numbers (widest for the text-only cells, n=193), and substantially wider for the small per-category cells, which we therefore do not interpret individually. Because every model answers the same items, we report paired McNemar tests and an item-level logistic regression in Appendix C, and release the per-item predictions. A single deterministic run gives a reproducible point estimate but not an estimate of run-to-run variance, which we discuss in the Limitations.

5 Results

5.1 Overall Performance

Table 1 reports overall accuracy, total runtime and average tokens per task on all 822 questions. `GPT-5.2` is clearly the strongest system at 0.787 accuracy, followed by `Gemini 2.5 Pro` at 0.628. All remaining models stay below 0.45, with `Llama 4 Scout` the weakest at 0.346, still well above the four-option random baseline of 0.25.

There is a visible *accuracy/resource-usage trade-off*: the two strongest systems also have by far the longest runtime and highest per-task token usage, while the cheaper variants (Mini, Flash, Scout) finish in 11–21 minutes but plateau around 0.35–0.43 (Table 6 in Appendix D breaks down dollar cost, which higher spend alone does not buy). Within each family, the heavier variant beats the lighter one (`GPT-5.2` > `GPT-5.4 Mini`; `Gemini Pro` > `Flash`; `Maverick` > `Scout`); pairwise McNemar tests confirm these differences are significant (Appendix C). Invalid or unparseable outputs (scored as wrong) are rare for most systems but reach 12.2% for `Gemini 2.5 Pro` and 8.2% for `Llama 4 Scout` (Appendix C), so part of their error count reflects formatting rather than reasoning failures.

5.2 Effect of Task Modality

Splitting the dataset by the three modality groups defined in Section 4 reveals the most consistent pattern in our results (Table 2): *every* model's accuracy drops monotonically from text to single-image to multi-image tasks. For `GPT-5.2` this is 0.974 $\rightarrow$ 0.831 $\rightarrow$ 0.632, and the same shape holds for the weakest systems. `GPT-5.2`'s text-only accuracy of 0.974 indicates that on purely linguistic Slovak reasoning the model is approaching the ceiling on this subset; the difficulty lies in multimodal integration.

The gap between models is also largest on text-only tasks (range 0.446–0.974, spread 0.53) and narrowest on multi-image tasks (range 0.293–0.632, spread 0.34): more visual content compresses systems toward a common, weaker level.

5.3 Performance by Category

Table 3 reports accuracy across all categories. Two patterns are visible. First, accuracy is *stable* across the six standard age-based categories (*Drobec* through *Senior*): for `GPT-5.2` the per-category range is only 0.733–0.800. The intended difficulty progression by age does not translate into a corresponding drop in model performance. Second, the text-only *Nevidiaci* categories are among the highest-scoring cells: *Nevidiaci ZŠ* is high for every model, and *Nevidiaci SŠ* is highest for the stronger models (though not for the weakest, e.g. `GPT-5.4 Mini` at 0.306). This is consistent with Section 5.2 and with the per-category image share in the bottom row of Table 3: the *Nevidiaci* categories are 0%

Model	Drobec	Bobrík	Benjamín	Kadet	Junior	Senior	Nev. ZŠ	Nev. SŠ	3.–4. r.	5.–7. r.	6.–9. r.
GPT-5.4 Mini	0.500	0.427	0.364	0.429	0.376	0.293	0.528	0.306	0.625	0.125	0.154
GPT-5.2	**0.778**	**0.733**	**0.799**	**0.800**	**0.765**	**0.738**	**0.917**	**0.972**	0.875	0.625	0.923
Gemini 2.5 Flash	0.500	0.467	0.416	0.457	0.465	0.354	0.639	0.361	0.500	0.375	0.385
Gemini 2.5 Pro	0.611	0.507	0.643	0.621	0.641	0.561	**0.917**	0.944	0.375	0.250	0.615
Llama 4 Maverick	0.333	0.427	0.422	0.386	0.382	0.415	0.583	0.444	0.250	0.500	0.154
Llama 4 Scout	0.278	0.387	0.331	0.386	0.341	0.280	0.472	0.417	0.250	0.250	0.385
Tasks (N)	18	75	154	140	170	164	36	36	8	8	13
Image-based (%)	89	96	84	88	75	81	0	0	100	100	92

Table 3: Accuracy across competition categories, with per-category task counts (N) and the share of image-based tasks in the bottom two rows. The two *Nevidiaci* categories (right of the vertical rule) contain only text-based tasks (0% image). The three shaded right-most columns are non-standard one-off categories with very small sample sizes (8, 8, and 13 tasks); results there should be read with caution. Given the small cells, per-category accuracies are best read as qualitative.

image-based, whereas the lower-scoring standard categories are 75–96% image-based. In our data, then, *accuracy varies more with input modality than with age-based category difficulty*.

The three rightmost shaded columns (*3.–4. roč.*, *5.–7. roč.*, *6.–9. roč.*) are non-standard categories used only in the special years *Karanténa* and *Prechádzky* and contain only 8, 8, and 13 tasks respectively. We include them for completeness but draw no conclusions from them.

5.4 Comparison with Student Performance

The *iBobor* platform publishes a per-year, per-category average student success rate, which we plot on the same 0–1 scale as model accuracy. This published aggregate reflects the platform's own scoring conventions (which can penalise incorrect answers), so we treat it as an approximate reference rather than a strict percent-correct figure. We compare model and student accuracy for three recent competition cycles (2018/2019, 2019/2020, 2020/2021) as a representative sample; Figure 2 shows the 2020/2021 results as a representative example, and the same qualitative picture holds in the other two years. A full per-year comparison is left to future work. `GPT-5.2` is usually at or above the average student in most categories; `Gemini 2.5 Pro` is competitive but less consistently so, sitting near the student average; the largest model-over-student gaps appear in the *Nevidiaci* categories where every task is textual. Smaller models (`GPT-5.4 Mini`, `Gemini 2.5 Flash`, `Llama 4 Maverick`, `Llama 4 Scout`) sit at or below the average student, often by a sizable margin. In short, on this multiple-choice subset, *frontier-tier LLMs already match or exceed the published average Slovak student accuracy in most categories, while mid-tier multimodal models do not*. We stress that this is a category-level comparison against published yearly aggregates rather than a task-matched one (see Limitations).

5.5 Token Usage Across Input Representations

Li et al. (2025) report that, for long-context retrieval, rendering text as an image can roughly halve the number of decoding tokens at little or no accuracy cost. We tested whether the same trade-off applies to short Bebras items by re-running, as single screenshots, the 612 tasks whose answer options are purely textual (otherwise screenshot answers cannot be automatically scored). This subset comprises 193 text-only, 308 single-image, and 111 multi-image tasks; the multi-image cell in Table 5 is therefore the 111-task textual-option slice, not the full 321 multi-image tasks of Table 2, and should not be generalised to all of them. Two further caveats apply to the *accuracy* comparison. The screenshot setting changes three things at once relative to the structured setting (the input representation, the required output, namely a free-text answer rather than a `choice_value`, and the scorer, a normalised text match rather than an exact value match), so any accuracy difference is confounded and cannot be attributed to the representation alone; the least confounded result here is the token-usage difference. Full results are in Table 5 (Appendix B).

The token-usage pattern is the *opposite* of what one would expect from the long-context setting: in almost every model/modality cell screenshot input *increases* token usage, often several-fold (e.g. `GPT-5.4 Mini` on text-only tasks: 358 → 1903 tok/q), though a few increases are small

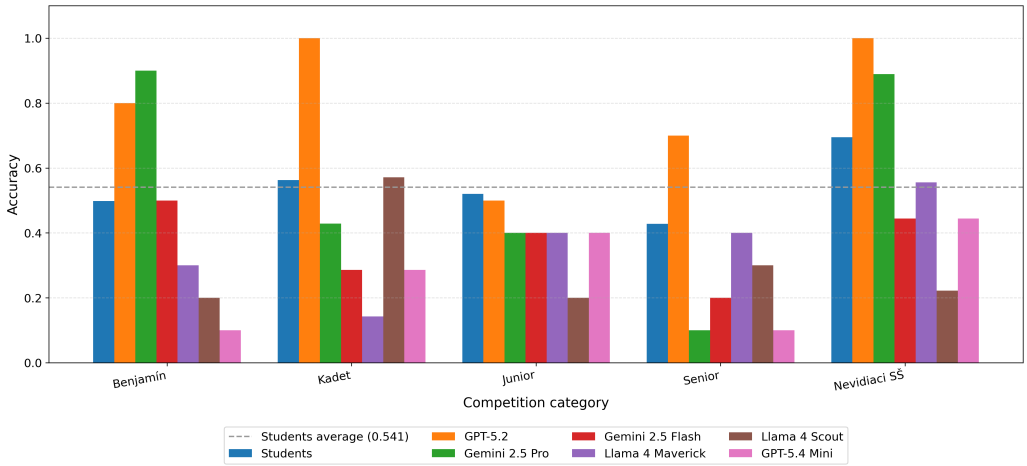

Figure 2: Accuracy of evaluated models versus average student results, competition year 2020/2021, per category. *Drobec* and *Bobrík* are omitted from this comparison because, in this year, their published student aggregates are dominated by canvas-based tasks that are excluded from our dataset.

(e.g. `Gemini 2.5 Flash` single-image: 1306 $\rightarrow$ 1324) and the sole decrease is `Gemini 2.5 Pro` on multi-image tasks (9526 $\rightarrow$ 7579). Accuracy effects, by contrast, are mixed in both sign and magnitude: screenshots are neutral-to-negative on text-only tasks and for the weaker models, but *improve* the strongest models on image tasks (`GPT-5.2`: 0.831 $\rightarrow$ 0.873 single-image, 0.705 $\rightarrow$ 0.820 multi-image; `Gemini 2.5 Pro` similarly). The clearest and least confounded finding is therefore the large token-usage increase: for these short tasks a screenshot's fixed visual-token overhead (set by image resolution, margins, and layout rather than text length) dominates any savings from replacing text, so the text-as-image trick appears to be a long-input rather than short-input optimisation (Appendix D).

5.6 Robustness Across Competition Years

Public, internet-accessible benchmarks risk contamination: a model may have seen its evaluation items during pretraining, inflating measured accuracy. The Slovak *iBobor* archive has been online since 2007, so the older year subsets are a realistic worst case for contamination, while later archived years are somewhat less exposed (though not necessarily clean relative to unknown model training cutoffs). As a sanity check we examined whether per-year accuracy trends *downward* for newer competition years, a contamination signature. On the text and single-image subset (complete for all six models), per-year accuracy correlates weakly and negatively with competition year for every model (Spearman ρ between -0.52 and -0.11), but the trend is not statistically significant for any model (smallest $p = 0.06$, `GPT-5.4 Mini`; Appendix C). This weak signal lies in the contamination direction yet is far from conclusive, and could equally reflect newer tasks being intrinsically harder. The figures illustrate a related but distinct point: Figure 4 (years 2018/2019 and 2019/2020, Appendix B), together with Figure 2, shows that the model-vs-student *ordering* (`GPT-5.2` $\gtrsim$ students $\gtrsim$ `Gemini Pro` $\gg$ smaller models) is stable across three competition cycles, with the *Nevidiaci* text-only categories remaining the easiest cells.

This is only *weak, indirect* evidence: training cutoffs are unknown, per-year cells are small and noisy, and properly ruling out memorisation would require dedicated detection (Golchin and Surdeanu, 2024; Sainz et al., 2023) or a held-out set of tasks published after model cutoffs. We therefore make no claim that our headline numbers are contamination-free.

6 Discussion

We draw three main conclusions. *First*, frontier-tier closed models (`GPT-5.2`, and partly `Gemini 2.5 Pro`) reach or exceed the published average Slovak student accuracy on the multiple-choice iBobor subset, while their lighter variants and open Llama 4 models do not. The smallest gap between the two ends of our model range appears on the hardest (multi-image) inputs, where every system struggles. *Second*, modality matters more than age-based category difficulty: across the six standard

age categories, accuracy varies far less than it does across the three modality groups, and the apparent "easiness" of the *Nevidiaci* categories is largely attributable to their being text-only (*Nevidiaci* is also a separate authoring track for visually impaired students, which the marginal tables cannot fully disentangle from modality). *Third*, short-task screenshot input does not save tokens; it costs them several-fold, with only mixed, model-dependent accuracy effects. These observations together suggest that current multimodal LLMs still treat text+image inputs as fundamentally more expensive and harder than text-only ones, even when the visual content is essential and the textual content is small.

Category accuracy tracks visual-content density. The category-level table (Table 3) is largely explained by how much visual content each category contains rather than by its nominal age difficulty. The two text-only *Nevidiaci* categories (0% image-based) are among the highest-scoring cells, clearly so for the stronger models, whereas the six standard age categories are uniformly image-heavy (75–96% image-based, bottom row of Table 3) and cluster at a lower, relatively flat accuracy. An item-level logistic regression of correctness on modality, category, and year (with model as a fixed effect) confirms this under controls: modality reduces deviance far more than category (likelihood-ratio $\chi^2(2) = 149$ vs. $\chi^2(7) = 43$), and the odds of a correct answer fall to 0.34 (single-image) and 0.28 (multi-image) relative to text-only, both highly significant (Appendix C). Category is a real but much weaker predictor. The practical implication is that any benchmark mixing modalities at the item level should report per-modality numbers, otherwise headline category accuracy mostly tracks visual-content density rather than reasoning depth, a confound that would also affect cross-language Bebras comparisons.

Implications for under-resourced languages. Our strongest model reaches a broadly comparable order of magnitude to Capdehourat et al. (2025)'s English Bebras results (≈70% for strong models), though the two settings are not directly comparable (26 image-as-text tasks in English vs. our 822 native multimodal Slovak tasks). The Slovak language is a plausible contributor (it is under-represented in pretraining), but our experimental setup does not isolate language effects from task design, prompt, model selection, dataset filtering, or single-run variance, so we treat this as a hypothesis rather than a conclusion. A clean test would require parallel Slovak and English versions of the same task set; constructing one is an attractive direction for future work, particularly given that the *Kaleidoscope* (Salazar et al., 2025) effort already covers 18 languages but not Slovak. More broadly, our results are consistent with the KALEIDOSCOPE finding that less-resourced languages are systematically under-served by closed-source frontier models; we contribute Slovak as one new data point rather than cross-language evidence of our own.

What this means for future Bebras-style evaluations. The token-as-image experiment (Section 5.5) suggests that reformulating short tasks visually is not a viable cost-reduction strategy, contrary to the long-context findings of Li et al. (2025). For short-form multiple-choice evaluation, structured text-plus-image input remains several-fold cheaper in tokens, with confounded accuracy differences that are comparable on text-only tasks and sometimes favour screenshots for the strongest models on image-task subsets. We therefore recommend that future Bebras-style benchmarks (i) report per-modality accuracy by default, (ii) preserve native task images rather than embedding them in screenshots, and (iii) include text-only tasks as a control group to disentangle visual reasoning from linguistic comprehension.

7 Conclusion

We introduce an 822-task multimodal Slovak Bebras dataset and use it to benchmark six modern LLMs across four dimensions: overall accuracy, task type, competition category, and against average student performance. The dominant effect is task modality: every model degrades monotonically from text to single-image to multi-image tasks, and category-level "difficulty" is largely a proxy for how much visual content a category contains. The best system clears average student level in most categories, but lighter variants and open models do not. Finally, contrary to recent long-context findings, rendering short iBobor tasks as screenshots costs several times more tokens while changing accuracy only inconsistently. Useful next steps include extending the dataset to canvas-based tasks, adding more recent open-weight models, evaluating chain-of-thought and few-shot prompting, and combining multiple short tasks into a single screenshot to test whether text-as-image becomes economical at higher input lengths.

Limitations

Single-run, single-prompt evaluation. Every task is evaluated exactly once per model with one fixed zero-shot prompt. This mirrors the actual contest, but it does not separate stable model ability from per-task variance, nor does it explore prompt sensitivity. Majority voting, repeated sampling, and a prompt sweep are left to future work.

Model versions and OpenRouter drift. All models were accessed through OpenRouter between January and March 2026; the underlying provider versions may change over time and across regions, limiting strict reproducibility. We report the snapshot we measured rather than a guarantee about future versions of the same model names.

Excluded task types. Canvas-based and Flash-based tasks (276 in total, 25% of those collected) are excluded because they cannot be reduced to a multiple-choice format with an automatically scorable single correct answer. The accuracy figures we report are therefore conditional on the multiple-choice subset.

Slovak language vs. task difficulty. Our experiment does not isolate the effect of the Slovak language from other factors (task design, modality, prompting). Differences relative to English-language Bebras results may or may not be language-driven, and we do not claim to have measured a clean "Slovak penalty".

Student comparison is not strictly matched. Student averages are taken from the iBobor platform's published yearly aggregates, which include all task types in a given year/category, whereas our models are scored only on the multiple-choice subset that survives filtering. This makes the comparison illustrative rather than strictly head-to-head.

Small-sample categories. The three non-standard categories (*3.–4.*, *5.–7.*, *6.–9. roč.*) contain only 8, 8, and 13 tasks respectively. We include them for completeness, but per-cell accuracy values there are not statistically meaningful.

Age category as a difficulty proxy. We use the age-based competition category as our notion of "nominal difficulty". Bebras additionally assigns each task an explicit difficulty level (easy/medium/hard), but this per-task label was not available through the public interface we scraped, so we could not analyse accuracy against it. Our claim is therefore specifically that, in our marginal breakdowns, accuracy varies little across age-based categories while showing a much larger association with modality; we have not adjusted for modality in an item-level model, nor do we claim accuracy is invariant to task difficulty in the Bebras sense.

Screenshot-experiment subset. The screenshot vs. structured comparison is restricted to the 612 tasks with purely textual answer options, since options that are themselves images cannot be matched to a text response in a unified scorer. This excludes a non-trivial slice of multi-image tasks (a known limitation of any automatically-scored screenshot setting).

Ethics Statement

The dataset is collected from the publicly accessible demo deployment of the Slovak *iBobor* platform with the explicit consent of the competition's national coordinator, who is acknowledged in the camera-ready version of this paper. Tasks are released to the public after each competition cycle ends; we did not scrape unreleased content. No personal data are collected: the student-performance figures we compare against are already published as aggregate yearly averages on the iBobor portal and contain no individual-level information. The task content itself is educator-authored competition material that, by the Bebras programme's design, contains no personal data and no offensive content; we additionally inspected the collected tasks and found none. Our use of the tasks, collected for research with the coordinator's consent, is consistent with their educational intent, and we ask that derivatives of the released dataset remain within research and educational contexts. Individual *iBobor* tasks are published under the Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0) licence; following the licensing and permission terms confirmed with the national coordinator, we will release the dataset under a compatible CC BY-SA 4.0 licence, with attribution to the original task authors and the Slovak *iBobor* organisers. The release will include the task JSON, images, screenshots, the evaluation prompts and scorer, and the raw model predictions and token counts.

Potential risks and misuses. We see limited risk in releasing already-public educational material. The intended use of the benchmark is to evaluate LLM reasoning on Slovak computational-thinking

tasks for research purposes; we discourage two foreseeable misuses: (i) training models on the released test set, since the dataset's value depends on it remaining a clean evaluation benchmark, and (ii) using model scores as an off-the-shelf substitute for student assessment in classrooms, since our results show that mid-tier models trail the average student by a wide margin and even the strongest model fails on about one in five tasks. The reported numbers describe model behaviour on a small set of zero-shot runs and should not be read as a claim that LLMs can replace human instructors of computational thinking.

Use of AI Assistants

AI-based assistants were used for language editing and to support literature search during the preparation of this paper. All research design, data collection, experiments, analysis, and scientific claims are the authors' own, and the authors take full responsibility for the content.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Uttama Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4232–4267. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Apify. Crawlee: a web scraping and browser automation library. https://crawlee.dev/. Accessed 2026.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 749–775. Association for Computational Linguistics.

Bebras International Challenge on Informatics and Computational Thinking. Bebras challenge. https://bebras.org. Accessed 2026.

Germán Capdehourat, María Eugenia Curi, and Víctor Koleszar. 2025. LLMs take on the Bebras challenge: How do machines compare to students? In *Proceedings of the 17th International Conference on Computer Supported Education – Volume 2: CSEDU*, pages 338–346. SciTePress.

Valentina Dagiene and Gerald Futschek. 2008. Bebras international contest on informatics and computer literacy: Criteria for good tasks. In *Informatics Education - Supporting Computational Thinking, Third International Conference on Informatics in Secondary Schools - Evolution and Perspectives, ISSEP 2008, Torun, Poland, July 1-4, 2008, Proceedings*, Lecture Notes in Computer Science, pages 19–30. Springer.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: multimodal large language models can see but not perceive. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, Lecture Notes in Computer Science, pages 148–166. Springer.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lei Huang, Weijiang Yu, Weitao Ma, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

iBobor. Informatický bobor. https://ibobor.sk/. Accessed 2026.

Yanhong Li, Zixuan Lan, and Jiawei Zhou. 2025. Text or pixels? it takes half: On the token efficiency of visual text inputs in multimodal LLMs. *arXiv preprint arXiv:2510.18279*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, Lecture Notes in Computer Science, pages 216–233. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

OpenRouter. OpenRouter: A unified interface for LLMs. https://openrouter.ai/. Accessed 2026.

Matús Pikuliak, Stefan Grivalsky, Martin Konopka, Miroslav Blsták, Martin Tamajka, Viktor Bachratý, Marián Simko, Pavol Balázik, Michal Trnka, and Filip Uhlárik. 2022. Slovakbert: Slovak masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Findings of ACL, pages 7156–7168. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 10776–10787. Association for Computational Linguistics.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, and 1 others. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. *arXiv preprint arXiv:2504.07072*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18761–18799. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15894–15939. Association for Computational Linguistics.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jeannette M. Wing. 2006. Computational thinking. *Commun. ACM*, 49(3):33–35.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Prompts and Scoring Details

For reproducibility we give the text-only prompt verbatim and summarise the multimodal and

screenshot adaptations and the scorer here; the full per-setting prompts, the scorer, and all raw predictions and token counts will be included in the released artifact.

Text-only system prompt. The system message for text-only tasks is, verbatim:

```
You will receive one
multiple-choice question.
The input may contain HTML
formatting such as bold text,
line breaks, lists, or tables.
Interpret this formatting as
part of the question.  Select
the correct option value (NOT
LABEL). Return JSON only in this
exact format:
{"id":  number, "choice_value":
"string"}
```

Multimodal prompt. For image tasks the message above is extended to state that images form part of the question and must be interpreted as such and, for multi-image tasks, that answer options may themselves contain images. Image placeholders in the task text are replaced, in order, by the corresponding images, passed as base64 data URLs.

Screenshot prompt. In the screenshot setting the entire task is supplied as a single image and the model is asked to return the exact answer text as written in the image, rather than a `choice_value`.

Scoring. In the structured settings the returned `choice_value` is matched exactly against the gold value; a returned label is mapped back to its value where unambiguous, and is otherwise logged as a parse error. In the screenshot setting the returned answer text is compared to the plain-text gold answer after lightweight normalisation; the exact normaliser will be included in the released artifact.

B Additional Tables and Figures

This appendix collects supporting material referenced from the main text: the full dataset statistics (Table 4), a text-only task example (Figure 3), the structured-vs-screenshot token-usage table (Table 5), and the per-year model-vs-student comparison for two further competition years (Figure 4).

C Item-level Statistical Analysis

The analyses here are computed from the released per-item predictions, joined to the gold answers and metadata. They cover the 4,142 model–item judgements over the eight standard age-based and *Nevidiaci* categories for which we have predictions: text-only and single-image tasks for all six models, and multi-image tasks for the four models whose per-item logs we retained.

Statistic	Value
Total collected tasks	1098
Excluded canvas-based tasks	153
Excluded Flash-based tasks	123
Final dataset	**822**
Text-only tasks	193
Image-based tasks	629
with one image	308
with multiple images	321
Competition years covered	16
Competition categories covered	11

Table 4: Main dataset statistics.

Modality versus category (controlled). We fit a logistic regression of answer correctness on input modality, competition category, and competition year, with model as a fixed effect. Likelihood-ratio tests show that modality contributes far more than category: adding modality to a model+category+year baseline gives $\chi^2(2) = 149.0$ ($p < 10^{-32}$), whereas adding category to a model+modality+year baseline gives $\chi^2(7) = 43.2$ ($p < 10^{-6}$). Relative to text-only tasks, the odds of a correct answer fall to 0.34 for single-image ($p < 10^{-23}$) and 0.28 for multi-image tasks ($p < 10^{-29}$). Category is therefore a real but substantially weaker predictor than modality, confirming the marginal pattern of Section 5 under controls.

Model-ranking significance. Pairwise McNemar tests on the 501 text and single-image items answered by all six models confirm the headline ranking: `GPT-5.2` > `Gemini 2.5 Pro` ($p \approx 1 \times 10^{-9}$), `Gemini 2.5 Pro` > `Gemini 2.5 Flash` ($p \approx 2 \times 10^{-21}$), and `GPT-5.2` > `GPT-5.4 Mini` ($p \approx 4 \times 10^{-57}$).

Invalid-output rates. The share of items returning an empty or unparseable answer (scored as wrong) varies markedly across systems: `Gemini 2.5 Pro` 12.2% and `Llama 4 Scout` 8.2%, versus $\leq 2.1\%$ for the others. For these two systems a non-trivial part of the measured errors are formatting failures rather than reasoning failures, which partly explains `Gemini 2.5 Pro`'s mod-

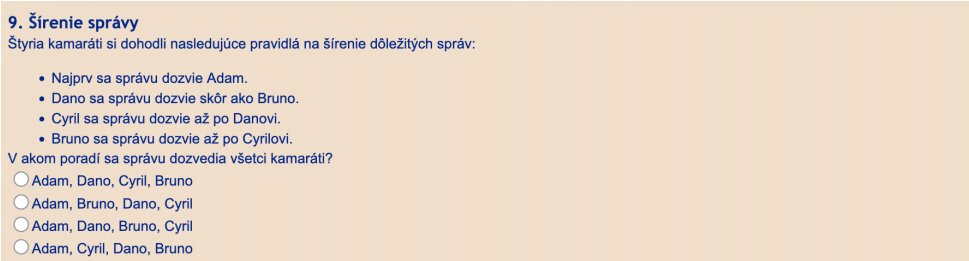

Figure 3: Example of a text-only iBobor task, *Šírenie správy* (*Spreading the News*), Nevidiaci ZŠ category, 2020/2021. The task is fully solvable from text; no visual content is required.

Model	Structured		Screenshot	
	Acc.	Tok/q	Acc.	Tok/q
Text-only tasks (n=193)				
GPT-5.4 Mini	0.523	358	0.389	1903
GPT-5.2	0.974	758	0.902	2740
Gemini 2.5 Flash	0.617	367	0.528	1790
Gemini 2.5 Pro	0.938	2636	0.907	3190
Llama 4 Maverick	0.591	355	0.373	2174
Llama 4 Scout	0.446	357	0.347	2177
Single-image tasks (n=308)				
GPT-5.4 Mini	0.364	494	0.354	2788
GPT-5.2	0.831	2297	0.873	4070
Gemini 2.5 Flash	0.419	1306	0.409	1324
Gemini 2.5 Pro	0.656	5972	0.731	6751
Llama 4 Maverick	0.367	743	0.299	2440
Llama 4 Scout	0.331	812	0.325	2440
Multi-image tasks (n=111)				
GPT-5.4 Mini	0.315	566	0.288	2911
GPT-5.2	0.705	4212	0.820	4678
Gemini 2.5 Flash	0.351	1017	0.387	1234
Gemini 2.5 Pro	0.477	9526	0.550	7579
Llama 4 Maverick	0.333	1055	0.306	2479
Llama 4 Scout	0.306	1223	0.261	2480

Table 5: Accuracy and tokens per task under **structured** input (text + inline images) versus **screenshot** input (the whole task rendered as a single image). Restricted to the 612 tasks with purely textual answer options so that screenshot answers can be scored unambiguously.

est accuracy despite its high token usage.

Per-year trend (contamination). On the text and single-image subset, per-year accuracy correlates weakly and *negatively* with competition year for all six models (Spearman ρ between -0.52 and -0.11), but the correlation is not significant for any model (smallest $p = 0.06$, `GPT-5.4 Mini`). The direction (newer years slightly harder) is the contamination signature we looked for in Section 5.6, but its weakness and lack of significance mean we can neither confirm nor exclude a small contamination effect; it could equally reflect newer tasks being intrinsically harder.

Model	Acc.	In	Out	$/1k
GPT-5.2	**0.787**	1.75	14.0	31.1
Gemini 2.5 Pro	0.628	1.25	10.0	57.6
Gemini 2.5 Flash	0.434	0.30	2.5	0.40
Llama 4 Maverick	0.408	0.15	0.60	0.14
GPT-5.4 Mini	0.374	0.75	4.5	0.43
Llama 4 Scout	0.346	0.08	0.30	0.08

Table 6: Estimated monetary cost. **In/Out** are OpenRouter list prices in USD per million prompt/completion tokens (retrieved May 2026); **$/1k** is the estimated cost per 1,000 tasks, from each model's mean tokens per task (Table 1) and its measured prompt/completion split.

D Cost

To convert token usage into money we combine each model's mean tokens per task (Table 1) with its measured prompt/completion split and the corresponding OpenRouter list prices; results are in Table 6. Prices were retrieved in May 2026 and may differ slightly from those in effect during the runs (Section 4), so these are estimates. The cost figures show three things. First, cost spans more than two orders of magnitude (from $0.08 per 1,000 tasks for `Llama 4 Scout` to $57.6 for `Gemini 2.5 Pro`), so the full 822-task evaluation costs under $0.10 on the cheapest model but about $26 on `GPT-5.2` and $47 on `Gemini 2.5 Pro`. Second, higher spend does not buy accuracy: `Gemini 2.5 Pro` is the costliest system (1.85× `GPT-5.2`) yet 16 accuracy points behind it. Third, this expense is dominated by *completion* (reasoning) tokens, which providers price several times higher than prompt tokens: completion accounts for ≈75% of `GPT-5.2`'s and ≈84% of `Gemini 2.5 Pro`'s output, versus ≤4% for the cheaper models, which emit little beyond the answer. Accuracy on *iBobor* thus comes largely from test-time reasoning, the dominant cost driver. (The completion share for `GPT-5.2` and `Gemini 2.5 Pro` is measured on the text and single-

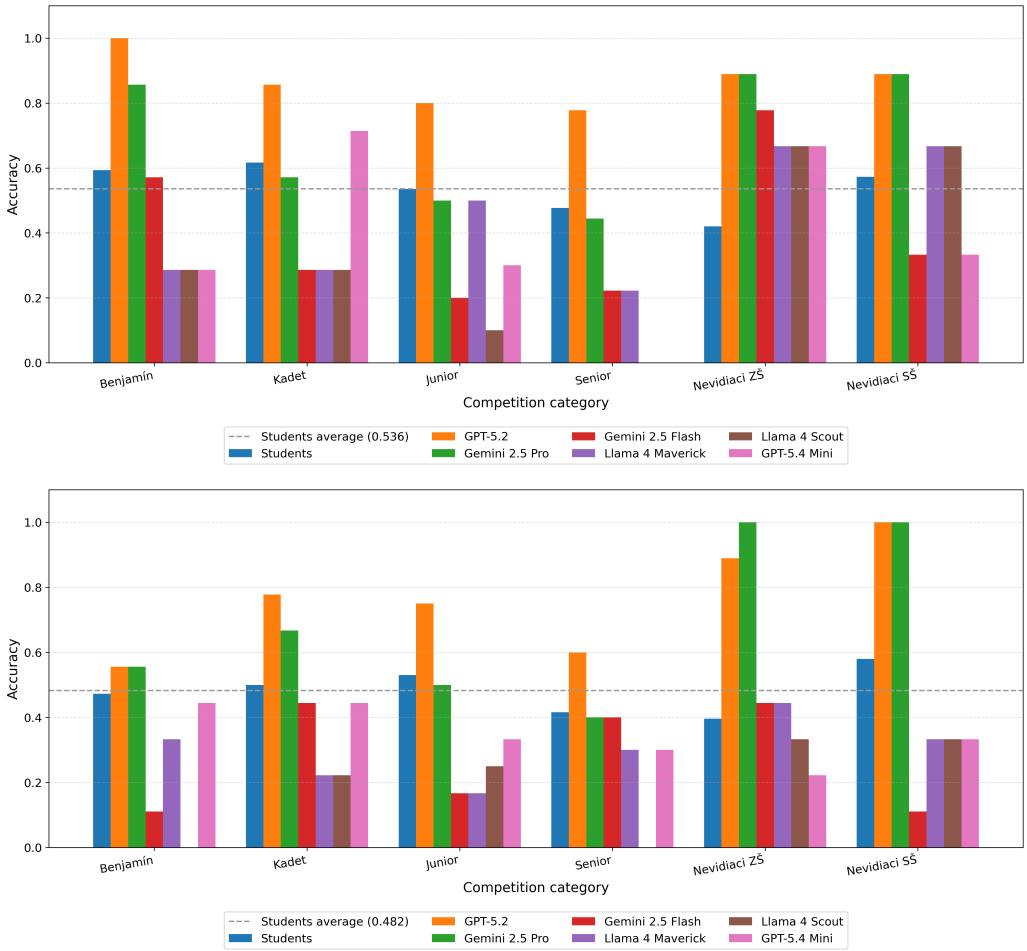

Figure 4: Model accuracy against average student results for two further competition years: 2018/2019 (top) and 2019/2020 (bottom). The qualitative picture matches Figure 2: `GPT-5.2` meets or exceeds the average student in most categories, `Gemini 2.5 Pro` is competitive, and smaller models trail. The full per-year model breakdown will be released with the dataset.

image subsets, whose per-item logs we retained, and assumed comparable on multi-image tasks.) In the screenshot setting (Section 5.5) the token increase is driven instead by a fixed visual-token overhead (image resolution, margins, and layout, not text length), so rendering these short tasks as images is uneconomical.